Journal of Statistical Planning and Inference 173 (2016) 47–63

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Exploring dependence between categorical variables: Benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms

Michail Papathomas [a,*], Sylvia Richardson [b]

[a] School of Mathematics and Statistics, University of St Andrews, The Observatory, Buchanan Gardens, St Andrews, KY16 9LZ, UK

[b] MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form 10 December 2015
Accepted 3 January 2016
Available online 15 January 2016

Keywords:
Bayesian model selection
Sparse contingency tables
Graphical models

ABSTRACT

This manuscript is concerned with relating two approaches that can be used to explore complex dependence structures between categorical variables, namely Bayesian partitioning of the covariate space incorporating a variable selection procedure that highlights the covariates that drive the clustering, and log-linear modelling with interaction terms. We derive theoretical results on this relation and discuss if they can be employed to assist log-linear model determination, demonstrating advantages and limitations with simulated and real data sets. The main advantage concerns sparse contingency tables. Inferences from clustering can potentially reduce the number of covariates considered and, subsequently, the number of competing log-linear models, making the exploration of the model space feasible. Variable selection within clustering can inform on marginal independence in general, thus allowing for a more efficient exploration of the log-linear model space. However, we show that the clustering structure is not informative on the existence of interactions in a consistent manner. This work is of interest to those who utilize log-linear models, as well as practitioners such as epidemiologists that use clustering models to reduce the dimensionality in the data and to reveal interesting patterns on how covariates combine.

1. Introduction

Detecting high-order interactions is becoming increasingly important for investigators in many fields of research. It is now understood that covariates may combine to affect the probability of an outcome, and that the effect of a particular covariate may only be important in the presence of other covariates. For example, in epidemiology it is of interest to examine the presence of interactions between smoking, environmental pollutants and dietary habits (Bingham and Riboli, 2004). In genetic association studies, it is of interest to detect gene–gene and gene-environment interactions in high dimensional data (Wakefield et al., 2010).

In this manuscript, we examine and discuss the relation between variable selection within Bayesian partitioning on one hand and log-linear modelling with interactions on the other, and the extend to which this relation can be explored in log-linear model search. Log-linear modelling is the most popular approach when searching for interactions, used by

* Corresponding author. Tel.: +44 1334461818.
E-mail address: M.Papathomas@st-andrews.ac.uk (M. Papathomas).

http://dx.doi.org/10.1016/j.jspi.2016.01.002

statisticians as well as practitioners in substantive applications. In a classical setting, attempting to fit a linear model with a large number of parameters sometimes requires an impractically large vector of observations to produce valid inferences (Burton et al., 2009). Within the Bayesian framework, the use of prior distributions alleviates identifiability or maximum likelihood estimation difficulties; see Dobra and Massam (2010). However, the space of competing models becomes vast, and model search algorithms like the Reversible Jump approach (Green, 1995) require a large number of iteration before they converge and produce reliable posterior model probabilities (Clyde and George, 2004; Dobra, 2009). With regard to contingency tables, the number of cells and possible graphical log-linear models that explain the cell counts increases exponentially with the number of covariates. For example, considering 20 covariates with 3 levels implies 3^{20} cells and approximately 1.5×10^{57} possible models.

Due to the difficulties associated with searching for interactions within a linear modelling framework, alternative approaches were adopted focusing on the reduction of the dimensionality in the data. Clustering is often the tool used to reduce dimensionality (see, for example Zhang et al., 2010), sometimes combined with a variable selection step (Chung and Dunson, 2009). Whilst log-linear modelling is a standard mathematical construction, there are many different clustering modelling approaches. For the purposes of this manuscript, we choose to focus on Bayesian clustering based on the Dirichlet process. The Dirichlet process produces flexible partitioning, allowing for the evaluation of the uncertainty with regard to the clustering of the subjects. We use a combination of Dirichlet process modelling and variable selection, implementing the modified variable selection step described in Papathomas et al. (2012), so that the covariates that contribute substantially to the clustering are identified.

We focus on categorical variables and log-linear models, as this is the standard framework for modelling interactions. In fact, for a set of categorical variables, where at least one is binary, there is a correspondence between log-linear and logistic regression modelling, and under certain conditions it is valid to translate inferences from the log-linear framework to the logistic one, regarding the presence of main effects and interactions; see Agresti (2002) and Papathomas (2015).

We explore the relation between log-linear modelling and clustering for two reasons. First, practitioners such as epidemiologists often use clustering in order to explore the manner in which covariates combine to affect the risk for disease; see Papathomas et al. (2011b). They frequently question if the clustering structures may inform in some way on the existence of interactions in associated log-linear models, and our investigation aims to provide some answers. Second, we aim to explore if any relation between log-linear modelling and clustering can be utilized to assist the exploration of large log-linear model spaces and the search for high-order interactions. The intuitive idea is that models that combine clustering and variable selection do not select covariates in accordance with the size of their marginal effect. Covariates are selected because they work together and combine with each other to create distinct groups of subjects. Consequently, this type of modelling may be able to inform on covariates that combine to describe the structure in the data, rather than covariates with a strong marginal signal.

In this manuscript, we are not concerned with the large-p problem, where thousands or hundreds of thousands of covariates are considered; see, for example, Hans et al. (2007), Richardson et al. (2010), or Cho and Fryzlewicz (2012) for a comprehensive review. Although our discussion is relevant to data sets of higher dimension, we focus on a relatively modest number of categorical variables, say one hundred or fewer, with fewer than twenty involved in interaction terms.

We demonstrate that inferences from clustering can potentially reduce the number of factors considered, by determining covariates that are independent of all others. Subsequently, the number of competing log-linear models is reduced, making the exploration of the model space feasible. This is crucial when analysing data that form large sparse contingency tables. We introduce a novel model search approach for a log-linear model space, informed by results from variable selection within clustering. We demonstrate that this model search algorithm can identify parts of the model space that contain models of low probability (thus helping to locate the highest probability model in less iterations, on average, compared to a less informed approach), especially in the presence of covariates that are independent of all other factors. With regard to limitations, first we show that there is no dependable correspondence between the covariate profile of the generated clusters and the log-linear model that best describes the data. More importantly, using simulated and real data, we show that variable selection within Bayesian clustering does not consistently detect marginal independence between covariates when the independent covariates form interaction terms with other factors.

Studies on the relation between the two different modelling approaches are not commonplace. In Dunson and Xing (2009), a Dirichlet process mixture of product multinomial distributions defines the prior on a set of categorical variables. Bhattacharya and Dunson (2012) model the joint distribution of categorical variables using simplex factor models. In contrast to our approach, variable selection switches are not considered in the aforementioned manuscripts, and no direct connection is made with log-linear model search. We are aware of three recent manuscripts that utilize clustering. The first is Marbac et al. (2014), where the clustering is applied to the covariates. This is different to the clustering we consider, widely used by practitioners, where the partitioning is applied to the subjects of the study. The second, Johndrow et al. (2014), has some connection to our work. In this preprint, the authors examine situations where the joint distribution implied by a sparse log-linear model has a low-rank tensor factorization. Relevant to our work is also the third, Zhou et al. (2015). This manuscript introduces and utilizes the idea that marginally independent variables reduce the dimensionality of the problem. This approach, central also to our work, was conceived and developed independently in parallel in our manuscript. The modelling in Zhou et al. (2015) with regard to marginal independence has similarities with the one we adopt, and significant differences. Our focus is different from Zhou as we utilize results from clustering to accelerate Bayesian log-linear graphical model selection with the Reversible Jump, a novel approach in log-linear model determination. We come back to these points

of comparison in the Discussion Section. Section 2, provides a brief description of the clustering and log-linear modelling approaches and contains concepts and notation important to the rest of the manuscript. In Section 3, we present theoretical results on the correspondence between marginal independence on one hand, and variable selection within the Dirichlet process clustering approach on the other, as well as a novel model search approach for log-linear models. Five simulated data sets are analysed in Section 4, and two real data sets in Section 5. We conclude with a discussion.

2. Clustering and log-linear models

2.1. A Dirichlet process clustering model

The Dirichlet process (DP) is especially suited to the problem of clustering observations $x_1, \ldots, x_n$, without pre-specifying the number of clusters. It is assumed that given parameters μ_i, x_i is drawn from $F(\mu_i)$. The mixing distribution over the parameters μ_i is denoted by G. A suitable prior for G is a Dirichlet process with scale parameter α and mean distribution G_0. Using G_0 and α, the DP partitions the μ_i parameters into a discrete set in a flexible way, allowing the sharing of information between different but similar observations. Dirichlet process mixture models have been thoroughly investigated in the past (Ferguson, 1973; Lo, 1984; MacEachern and Müller, 1998; Walker et al., 1999; Green and Richardson, 2001). They are used in a wide range of applications, including epidemiology and genetic studies (Huelsenbeck and Andolfatto, 2007; Dunson et al., 2008; Sinha et al., 2010; Reich and Bondell, 2011).

We adopt the conjugate Dirichlet process mixture model used in Molitor et al. (2010) and Papathomas et al. (2011b) for profiling patterns of covariates in epidemiological studies. For subject i, a covariate profile x_i is a vector of categorical covariate values $x_i = (x_{i1}, \ldots, x_{iP})$, where P is the number of covariates. Let $z = \{z_1, \ldots, z_n\}$, where z_i is an allocation variable, so that $z_i = c$ denotes that subject, i, belongs to cluster c. Denote with $\phi_p^c(x)$ the probability that the pth covariate $x_{\cdot p}$ is equal to x, when the individual belongs to cluster c. Given that $z_i = c$, covariate $x_{\cdot p}$ has a multinomial distribution with cluster specific parameters $\phi_p^c = [\phi_p^c(1), \ldots, \phi_p^c(M_p)]$. Here, M_p denotes the number of categories of $x_{\cdot p}$. We assume that, a priori, $\phi_p^c \sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_{M_p})$, Denote with $\psi = \{\psi_c,\ c \in N\}$ the probabilities that a subject is assigned to cluster c. We adopt a flexible 'stick-breaking' prior on the allocation weights ψ_c, with a random parameter α (West, 1992; Ishwaran and James, 2001). For $\phi = \{\phi_p^c,\ c \in N, p = 1, \ldots, P\}$, the model is written as,

$$
\begin{aligned}
&x_i|z, \phi \sim \prod_{p=1}^{P} \phi_p^{z_i}(x_{ip}) \quad \text{for } i = 1, 2, \ldots, n.\\
&\phi_p^c(x_{ip}) \sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_{M_p}) \quad \text{for } c = 1, 2, \ldots\\
&P(z_i = c|\psi) = \psi_c \quad \text{for } i = 1, 2, \ldots, n, \text{ and } c = 1, 2, \ldots\\
&\psi_c = V_c \prod_{l<c}(1 - V_l) \quad \text{for } c = 2, 3, \ldots \text{ with } \psi_1 = V_1,\\
&V_c \sim \text{Beta}(1, \alpha) \quad \text{for } c = 1, 2, \ldots.
\end{aligned}
$$

This implies the more recognizable mixture for the likelihood of the covariate observations,

$$Pr(x_i|\phi, \psi) = \sum_{c=1}^{\infty} \Pr(z_i = c|\psi) \prod_{p=1}^{P} \Pr(x_{ip}|z_i = c) = \sum_{c=1}^{\infty} \psi_c \prod_{p=1}^{P} \phi_p^c(x_{ip}).$$

To identify the covariates that are important for the formation of clusters we consider the variable selection approach described in Papathomas et al. (2012), which is inspired from Chung and Dunson (2009). In summary, consider cluster specific binary indicators, γ_p^c, so that $\gamma_p^c = 1$ when covariate $x_{\cdot p}$ is important for allocating subjects to cluster c; otherwise $\gamma_p^c = 0$. Denote by $\pi_p(x_{ip})$ the marginal probability that covariate $x_{\cdot p}$ takes the value x_{ip}, $P(x_{\cdot p} = x_{ip})$. Note that caution should be exercised when interpreting this probability, as it is linked to the sampling frame. The probability that covariate $x_{\cdot p}$ is observed as x_{ip}, when subject, i, belongs to cluster c, is written as,

$$P(x_{\cdot p} = x_{ip} \mid z_i = c) = [\phi_p^c(x_{ip})]^{\gamma_p^c} \times [\pi_p(x_{ip})]^{(1-\gamma_p^c)}. \tag{1}$$

Utilizing $\pi_p(x_{ip})$ in (1) when $x_{\cdot p}$ does not contribute to subject allocation to cluster c is intuitively appropriate, as $P(x_{\cdot p} = x|z_i = c) = P(x_{\cdot p} = x)$ implies by Bayes Theorem that $P(z_i = c|x_{\cdot p} = x) = P(z_i = c)$. Now, we can write,

$$\pi_p(x_{ip}) = P(x_{\cdot p} = x_{ip}) = \sum_c \psi_c [\phi_p^c(x_{ip})]^{\gamma_p^c} \times [\pi_p(x_{ip})]^{(1-\gamma_p^c)}.$$

We assume that the γ_p^c are independent Bernoulli variables with $\gamma_p^c \sim \text{Bernoulli}(\rho_p)$, $0 < \rho_p < 1$. Here, ρ_p describes the probability that covariate $x_{\cdot p}$ is important for the partitioning of the subjects, in relation to the whole process rather than a specific cluster. For ρ_p, we consider a sparsity inducing prior with an atom at zero, so that $\rho_p \sim 1_{\{w_p=0\}}\delta_0(\rho_p) + 1_{\{w_p=1\}}\text{Beta}(\alpha_\rho, \beta_\rho)$, where $w_p \sim \text{Bernoulli}(0.5)$. This prior is appropriate when it is required to clearly discriminate

Table 1a
Cluster profiles in hypothetical simple illustration, defined by the ϕ_p^c multinomial probabilities, for covariate $x_{.p}$ and cluster c.

	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	$x_{.5}$	$x_{.6}$
Cluster 1	(0.01, 0.3, 0.69)	(0.01, 0.3, 0.69)	(0.1, 0.1, 0.8)	(0.1, 0.1, 0.8)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)
Cluster 2	(0.01, 0.5, 0.49)	(0.01, 0.5, 0.49)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)
Cluster 3	(0.29, 0.7, 0.01)	(0.29, 0.7, 0.01)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)	(0.8, 0.1, 0.1)

Table 1b
Summary cluster profiles in hypothetical simple illustration. The '<' ('>') symbol denotes that observation x of covariate $x_{.p}$ in cluster c is more (less) likely compared to the average in the whole sample; otherwise, the '0' symbol is used.

	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	$x_{.5}$	$x_{.6}$
Median(ρ_p)	0.8	0.8	0.9	0.9	0.001	0.001
Cluster 1	<<>	<<>	< 0 >	< 0 >	000	000
Cluster 2	< 0 >	< 0 >	> 0 <	> 0 <	000	000
Cluster 3	>><	>><	> 0 <	> 0 <	000	000

between important and non-important covariates. The Dirichlet process model described in this Section is fitted using the R package PReMiuM (Liverani et al., 2015).

To create an easily interpretable clustering end-product, whilst the rich MCMC output is utilized and uncertainty is accounted for, we have adopted the model averaging approach described in Papathomas et al. (2012). One aspect of this approach is the derivation of a specific partition that best represents the variable clustering of the subjects during the MCMC run. We refer to this as the 'representative partition'. To clarify our model and notation we give a simple illustrative example. Consider six categorical covariates, $x_{.1}, \ldots, x_{.6}$, taking values 0,1 and 2. Suppose that subjects are typically allocated into three sub-populations, with probabilities $\psi_1 = 0.3$, $\psi_2 = 0.3$ and $\psi_3 = 0.4$. The multinomial probabilities for the six covariates, given the allocation z_i of subject i, is given in Table 1a. For instance, for $z_2 = 3$ the second subject is allocated to the third group, and the multinomial probabilities for $x_{.1}$ with regard to that subject are, $\phi_1^3 = (0.29, 0.7, 0.01)$. Covariates $x_{.5}$ and $x_{.6}$ clearly do not contribute to the clustering of the subjects, as the multinomial probabilities are the same across clusters. This implies that $\gamma_5^c = \gamma_6^c = 0$ for all c. The proportions for the covariate values across the whole sample can be evaluated in accordance with the ψ_c and ϕ_p^c parameters. For example, $\pi_1(0) = 0.3 \times 0.01 + 0.3 \times 0.01 + 0.4 \times 0.29 = 0.122$. For $x_{.5}$ and $x_{.6}$ this evaluation is trivial; for example $\pi_5(0) = 0.8$. After sampling from this population, a hypothetical summary profile of the three clusters can be derived using the posterior distributions of the model parameters; see Table 1b. For each covariate $x_{.p}$ and each possible observation $x = 0, 1, 2$, we consider the 95% credible interval (CI) for the difference between the probability $\phi_p^c(x)$ of attribute x in group c, and the corresponding frequency of $x_{.p} = x$ in the whole sample. Suppose that, with regard to the first group and the first covariate, the two CIs that correspond to $x = 0, 1$ are both below zero, whilst the CI that corresponds to $x = 2$ is above zero. So, for subjects in the first group, it is less likely to observe 0 or 1 at the first covariate, compared to the whole sample, and more likely to observe 2. We denote this information with the '<' and '>' symbols. We use the '0' symbol when the CI contains zero. In Table 1b, where we also provide hypothetical posterior medians for the selection probabilities ρ_p, $p = 1, \ldots, 8$, one can see the hypothetical summary structure in the population.

2.2. Log-linear graphical models

Denote with $\mathcal{P}$ the finite set of the P categorical covariates or factors. The resulting data can be arranged as counts in a P-way contingency table. A Poisson log-linear interaction model is a generalized linear model where the data are the cell counts of the contingency table; see Supplemental material, Section S1 (Appendix B), for a formal definition of an interaction term in a log-linear model. The number of all possible log-linear models is $2^{(2^P)}$. It can be very large for non-trivial applications. For example, the number of possible log-linear models for six factors is approximately 184×10^{19}. Graphical models are a subset of the class of log-linear models. They are represented by a graph where each node (or vertex) is an element of $\mathcal{P}$. Any two nodes may be connected by an edge. Nodes not connected directly by a single edge are independent conditionally on the factors represented by all other nodes (pairwise Markov property). Also, conditionally on nodes to which $x_{.p}$ is directly connected, $x_{.p}$ is independent of all other nodes (local Markov property). Finally, two sets of nodes are independent when they are separated by another set, conditionally on the separating set (global Markov property); see Lauritzen (2011) for more details. The number of possible graphical models is 2^H, where $H = P!/(2(P-2)!)$, assuming the intercept and all factor main effects are included in the model. For example, the number of possible graphical models for six covariates is 32768.

3. Results on marginal independence and a novel model search algorithm

3.1. Clustering and independence

Theorem 1. *Consider random variables $x_{.p}$ and $x_{.q}$, $1 \le p, q \le P$, $p \ne q$. If $\sum_{c=1}^{C} \gamma_p^c \times \gamma_q^c = 0$ then $x_{.p}$ and $x_{.q}$ are independent.*

Proof. See Appendix A.

Theorem 2. *Consider a set of random variables* $\{x_{\cdot 1}, \ldots, x_{\cdot P}\}$. *If, for some* $p \in \{1, \ldots, P\}$, $\sum_{c=1}^{C} \gamma_p^c \times \gamma_q^c = 0$, *for all* $q \neq p$, *then* $x_{\cdot p}$ *is independent of* $\{x_{\cdot 1}, \ldots, x_{\cdot P}\} \setminus x_{\cdot p}$.

Proof. See Appendix A.

Note that pairwise independence does not imply independence between sets of random variables. For example, if $x_{\cdot 1}$ is independent of $x_{\cdot 2}$ and of $x_{\cdot 3}$, it is not implied that $x_{\cdot 1}$ is independent of $\{x_{\cdot 2}, x_{\cdot 3}\}$. It is also crucial to note that the converse of Theorems 1 and 2 is not necessarily true. The previous Theorems lead to the following Corollary,

Corollary. *Consider a set of random variables* $\{x_{\cdot 1}, \ldots, x_{\cdot P}\}$. *If for some* $p \in \{1, \ldots, P\}$, $\sum_{c=1}^{C} \gamma_p^c = 0$, *then* $x_{\cdot p}$ *is independent of* $\{x_{\cdot 1}, \ldots, x_{\cdot P}\} \setminus x_{\cdot p}$.

Therefore, if the selection probability ρ_p for $x_{\cdot p}$ is zero or close to zero, something that implies that $\sum_{c=1}^{C} \gamma_p^c$ is also zero or close to zero, we can assume that $x_{\cdot p}$ is not connected with an edge with another covariate. If our interest lies in exploring interactions, to reduce the dimensionality of the problem when fitting log-linear models to sparse contingency tables, $x_{\cdot p}$ could be removed from the analysis.

3.2. Construction and interpretation of matrix $\boldsymbol{T}_\gamma$

Considering the results in Section 3.1, we construct $\boldsymbol{T}_\gamma$, a matrix that summarizes the variable selection output, and translates it into information that is relevant to log-linear modelling. The algorithm for the formation of $\boldsymbol{T}_\gamma$ is given below.

- For iteration i_t and for each cluster c with more than one subject, form matrix $\boldsymbol{T}^{c,i_t}$, so that element (p_1, p_2), $1 \leq p_1 < p_2 \leq P$ is either zero or one, and equal to $\gamma_{p_1}^c(i_t) \times \gamma_{p_2}^c(i_t)$. All other matrix cells are empty.
- Sum up all matrices $\boldsymbol{T}^{c,i_t}$, weighing by cluster size, to create an information matrix $\boldsymbol{T}_\gamma$,

$$\boldsymbol{T}_\gamma = \sum_{i_t} \sum_{c} n_{c,i_t} \times \boldsymbol{T}^{c,i_t}$$

  where n_{c,i_t} is the size of cluster c at iteration i_t. Therefore, $\boldsymbol{T}_\gamma$ is a straightforward summary of all $\boldsymbol{T}^{c,i_t}$ matrices into one, with small clusters contributing less to this summary.
- For ease of interpretation reweight the elements of $\boldsymbol{T}_\gamma$ so that the maximum element is one, $\boldsymbol{T}_\gamma = (max\{\boldsymbol{T}_\gamma\})^{-1} \times \boldsymbol{T}_\gamma$.

Matrix $\boldsymbol{T}_\gamma$ is constructed in such a manner so that if element $t_\gamma(p_1, p_2)$, $1 \leq p_1 < p_2 \leq P$, is close to zero, this implies that an edge between $x_{\cdot p_1}$ and $x_{\cdot p_2}$ is not likely to be present in a highly supported graphical model.

3.3. A modified log-linear model search algorithm

In this subsection, we propose a novel model comparison approach based on the Reversible Jump MCMC algorithm implemented in Papathomas et al. (2011a). We allow for the removal, addition or replacement of one edge in the graph with another. Whilst in the aforementioned manuscript the choice of edge was completely random, we now *inform this choice* by the clustering output using $\boldsymbol{T}_\gamma$.

To propose the addition of an edge to the currently accepted model, we consider the elements of $\boldsymbol{T}_\gamma$ that correspond to pairs of covariates not currently connected with an edge, transform so that they sum to one, and sample an edge using the derived probabilities. To suggest an edge for removal, we consider the elements of $\boldsymbol{T}_\gamma$ that correspond to pairs of covariates already connected with an edge, transform so that they sum to one, and sample an edge using complimentary probabilities. To choose one edge to replace another, we sample both edges as previously. A detailed demonstration of the calculations described in this subsection is presented in the Supplemental material, Sections S2 and S3 (Appendix B).

4. Simulation studies

The translation we implement between clustering and log-linear model search is novel. We therefore present an extensive range of simulation studies to demonstrate advantages and limitations. The first describes a relatively simple dependence structure. More complex structures are studied in the next two simulations, whilst the last two demonstrate the benefit of our approach with regard to the analysis of sparse contingency tables.

4.1. The simulated data sets

The specifications for the five simulations are shown in Table 2. For simulations 1–3, the majority of the subject observations (80%) is simulated using Model 1. The rest of the subjects are simulated using Models 2 and 3 in a balanced

Table 2
Simulation specifications.

	Number of subjects	Number of covariates	Number of levels of covariates	Number of cells in contingency table	Approximate number of models	Number of covariates that form interactions
Simulation 1	10,000	10	2	1024	3.5184×10^{13}	7
Simulation 2	10,000	10	2	1024	3.5184×10^{13}	6
Simulation 3	10,000	10	2	1024	3.5184×10^{13}	9
Simulation 4	5,000	20	3	3.4×10^{9}	1.5×10^{57}	6
Simulation 5	10,000	100	2	1.27×10^{30}	2^{4950}	8

Table 3
MCMC specifications for the clustering analyses, and also for the log-linear model comparison Reversible jump chains. Clustering analyses were performed using the R package PReMiuM. Reversible jump analyses were performed using Matlab code. All analyses performed on a PC equipped with an Intel(R) Core(TM)i7-2600K CPU 3.40 GHz with 8GB RAM.

Clustering algorithms	Burn-in	Iterations after burn-in	Run time in minutes (approx.)	Comment
Simulation 1	40,000	20,000	24	
Simulation 2	40,000	20,000	24	
Simulation 3	40,000	20,000	24	
Simulation 4	100,000	20,000	30	
Simulation 5	100,000	20,000	90	
Edwards and Havranek data (CHD)	40,000	20,000	3	
Genetic-environmental data	40,000	20,000	10	
Reversible jump chains	**Burn-in**	**Iterations**	**Run time in minutes**	**Comment**
Simulation 1	10,000	100,000	420	
Simulation 2	10,000	100,000	420	
Simulation 3	10,000	100,000	420	
Simulation 4	2,000	10,000	360	after discarding 14 covariates
Simulation 5	50,000	10^6	240	after discarding 92 covariates
Edwards and Havranek data (CHD)	20,000	10^6	65	
Genetic-environmental data	20,000	10^6	65	after discarding 18 SNPs

manner. The models are presented in Fig. 1. Simulation 1 is based on two distinct sets of covariates, where covariates that belong to different sets are independent. Simulations 2 and 3 describe more complex structures compared to simulation 1, since interaction terms share common covariates. We provide additional information on the design matrices and parameter coefficients of the utilized log-linear models in the Supplemental material, Section S4 (Appendix B). We used three models to generate each simulated data set, rather than one, in order to emulate more accurately the variability and complexity within a real data set.

Two more simulated data sets were created to demonstrate how our approach can be used for the analysis of sparse contingency tables. In simulation 4, only six out of twenty factors are important for explaining the variability associated with the cell counts. In simulation 5, only eight out of 100 factors are important for explaining the variability associated with the cell counts. Three models were used for the generation of the fourth and fifth simulated data sets, seen in Fig. 1, with probabilities {032%, 0.29%, 0.29%} and {0.8%, 0.1%, 0.1%} respectively.

The size of the model space in simulations 4 and 5 renders conventional model comparison algorithms like the reversible jump MCMC unfeasible. The cluster specific variable selection approach should detect that 14 and 92 covariates respectively are not important. This will allow for the removal of these covariates from subsequent analyses, forming a drastically smaller model space that can be explored in practice.

4.2. MCMC specifications, prior distributions and model search strategies

Information on the size of the chains, as well as run times, is provided in Table 3. The log-linear models were fitted and compared within the reversible jump MCMC framework described in Papathomas et al. (2011a). Simulation 4 contains factors with three levels each. Subsequently, models contain, on average, a larger number of parameters compared to the other simulations, resulting in a slower Reversible Jump algorithm. Hence, the relatively small number of iterations. Samples are rather small for accurately estimating posterior probabilities of less prominent models, in model spaces as large as the ones we consider. However, these chains provide valuable information for the mixing performance of the different reversible jump MCMC algorithms.

The following prior specifications were adopted. For the clustering Dirichlet process model we considered a sparse prior for ρ_p with a point mass at zero (see Section 2.1), to force a clear distinction between the covariates that contribute to the clustering and the ones that do not. Conjugate Dirichlet priors with $\lambda_1 = \cdots = \lambda_{M_p} = 0.5$ were adopted for the ϕ_p^c parameters. Chains were initialized by allocating subjects randomly to ten groups. Initial values for all other model

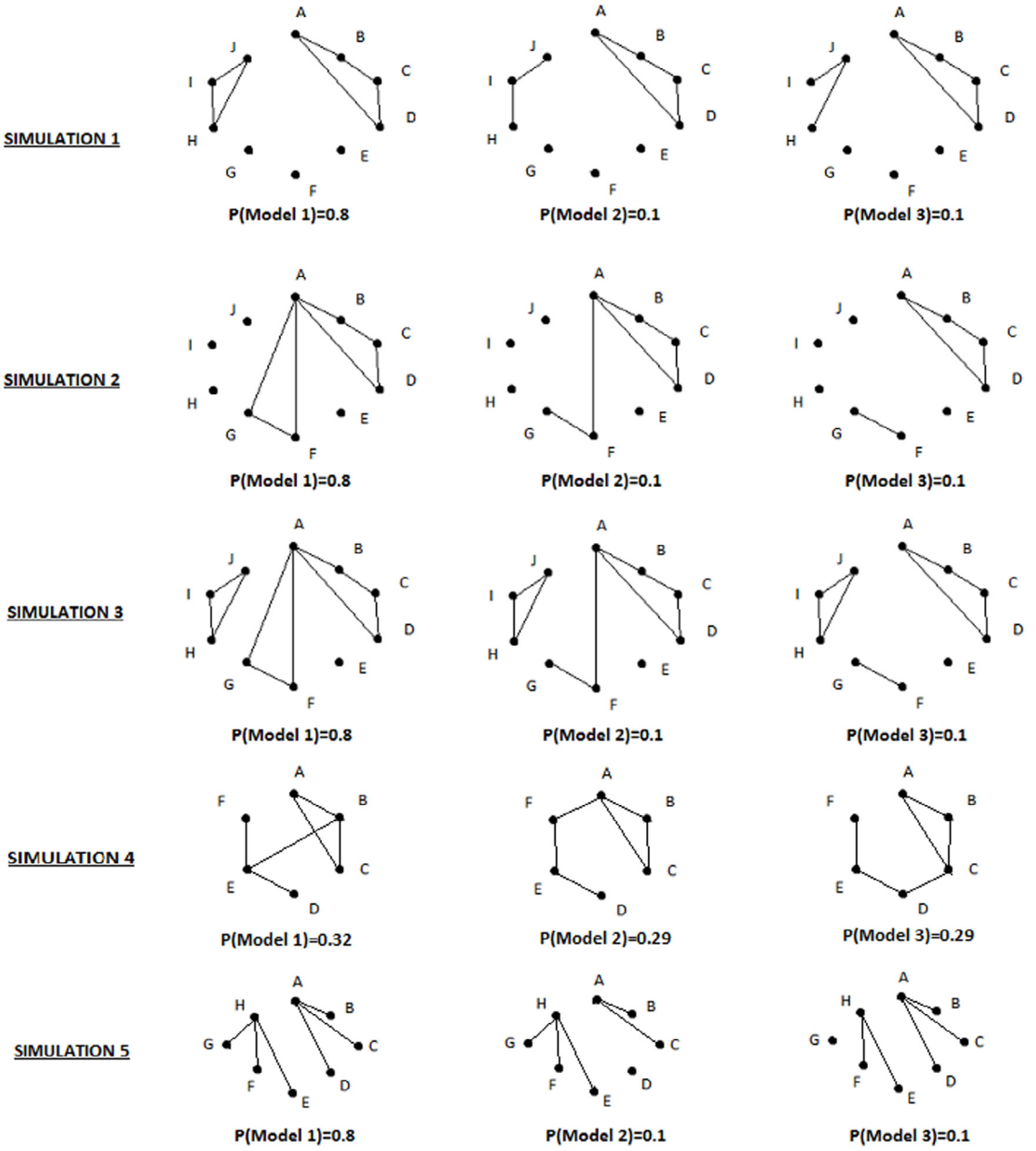

Fig. 1. The graphical models used for the five simulations.

parameters were random. Regarding the log-linear model comparison analyses, unit information priors (Ntzoufras et al., 2003) were adopted for the model parameters. All graphs are equally likely a priori. The majority of the specifications described above are also adopted in the real data analyses presented in Section 5, with differences indicated clearly therein.

Following standard practice when building a reversible jump MCMC chain, in 60% of the iterations, a new set of values for the parameters of the currently accepted model is proposed. A jump to a different graphical model is attempted in 40% of the iterations, where it is equally likely to attempt the addition, removal or replacement of one edge with another. We compare four model search strategies:

(a) Uniformly random selection. An unrefined model search strategy where all candidate edges are equally likely to be selected.
(b) The cluster specific approach described in Section 3.3.
(c) A combination of (a) and (b), where (a) is employed in 30% of the iterations and (b) in 10% of the iterations.
(d) A balanced combination of (a) and (b) where the two model search approaches are each employed in 20% of the iterations.

In all analyses, proposals for the model parameters are derived as in Papathomas et al. (2011a) manuscript where the unrefined model search strategy (a) is adopted. To allow for an intelligible comparison with this standard approach, we refer to the Reversible jump algorithm that employs (a) as the PDV approach using the authors' initials. We do not refer to (a) as PDV when covariates are discarded after implementing the clustering algorithm, because this is not a standard step. Note that parameter proposals could also be constructed following Forster et al. (2012), although the two approaches share many characteristics.

4.3. Simulation results

4.3.1. Variable selection within clustering and marginal independence

The flexible clustering algorithm discriminated clearly between important and unimportant covariates in all five simulations; see Table 4 for the posterior median selection probabilities ρ_p. Regarding simulations 4 and 5, the original model space contains 1.5×10^{57} and 2^{4950} graphical models respectively. Implementing the PDV algorithm on such vast model spaces is not feasible, since model comparison would be compromised in terms of convergence and numerical stability. For simulation 4, the variable selection approach described in Section 2.1 correctly reduced the number of covariates to six, after discarding 14 covariates with posterior median selection probabilities less than 0.14, whilst $E(\rho_p) < 0.0045$, $p = 7, \ldots, 20$. Regarding simulation 5, the number of covariates was correctly reduced to eight, with posterior median selection probabilities for the 92 unimportant covariates equal to zero or less than 0.01.

4.3.2. The representative cluster profiles in relation to the presence of interactions

In most simulations we observe some correspondence between the observed clustering structure and the simulated interactions. However, this correspondence is often blurred, and it is not obvious how to infer and untangle the different interaction terms simply by inspecting the cluster profiles shown in Table 4. For simulation 1, three clusters were highlighted in the output summary, as indicated by the patterns of '>' and '<' (see the end of Section 2.1). Clusters 1 and 2 correspond to the simulated 'ABCD' and 'HIJ' interactions. The posterior median for the selection probability for 'C' is only slightly lower than the medians of other important covariates, however 'C' does not appear to contribute to the formation of the cluster profiles as strongly as the other important covariates. Cluster 3 clearly corresponds to the 'HIJ' interaction. In accordance to the simulation set-up, 'E' and 'F' have very low selection probabilities. Hence in this simulation, the cluster profile 'matches' quite clearly the simulated interactions. For simulation 2, five clusters were highlighted in the output. Clusters 2–5 seem to correspond to the 'ABCD' and 'AFG' interactions, and the selection probabilities for 'H','I' and 'J' are low in accordance with the simulation set-up. Two clusters were highlighted in simulation 3. Their profiles seem to correspond to the 'ABCD' and 'AFG' interactions. The posterior selection probabilities for 'H', 'I' and 'J' are as high as the posterior medians of the other important covariates while that of 'E' is small, in accordance with the simulation mechanism. With regard to the fourth simulated data set, Table 4 presents results from the flexible clustering analysis in relation to the first six covariates, correctly selected by the clustering algorithm. Six clusters comprise the representative partition, but do not display clear separating patterns suggestive of the existence of specific interactions. This is also the case for simulation 5.

Overall we see that, although suggestive in some cases, the covariate profiles of representative clusters do not inform conclusively on interaction terms within a log-linear modelling framework. This note of caution is of interest to practitioners that employ clustering approaches, as the relation between covariate profiles and interactions within a linear modelling framework is often a matter of inquiry.

4.3.3. The derived $\boldsymbol{T}_\gamma$ matrices

The constructed $\boldsymbol{T}_\gamma$ matrices are shown below. We display with bold font the values of elements that correspond to an existing edge in the most probable model; see Section 4.3.4 for posterior model probabilities.

The $\boldsymbol{T}_\gamma$ matrices recover the graph of the most likely model well for Simulations 1 and 2, as expected from our discussion of the representative profiles. In terms of picking up existing or non-existing edges, it is clear in simulations 1–3 that, overall, smaller weight is given to non-existing edges, compared to existing ones. We also notice a 'spill-over' effect in the $\boldsymbol{T}_\gamma$ matrices, with blocks of high valued elements corresponding to important covariates that are not connected in the simulated graph.

In simulations 4 and 5, considering the important covariates, the elements of $\boldsymbol{T}_\gamma$ are all large, whether they correspond to an existing edge or not. This illustrates that the converse of the Theorems in Section 3.1 does not hold. There is no significant difference in the derived $\boldsymbol{T}_\gamma$ matrices, when the clustering is performed again on the reduced set of covariates.

Importantly, small elements in the $\boldsymbol{T}_\gamma$ matrices *always correspond to a non-existing edge*. They never indicate that an existing edge is absent, something that would be detrimental to a model search algorithm. If the value of an element $t_\gamma(p_1, p_2)$ is low, say less than 0.1, then it is always the case that the edge between x_{p_1} and x_{p_2} is absent from the high probability graphical model. Elements t_γ that correspond to existing edges are usually much larger, at least one or two orders of magnitude larger compared to elements with a clearly low value. These results confirm the correspondence between the two types of structures, the specificity of the pattern of small elements in $\boldsymbol{T}_\gamma$, and highlight the potential role of clustering

Table 4
Cluster profiles for the five simulations. In parenthesis the number of subjects typically allocated to each representative cluster. All posterior median selection probabilities for the remaining 14 covariates in Simulation 4 were less than 0.14. Posterior median selection probabilities for the remaining 92 covariates in Simulation 5 were either equal to zero or smaller than 0.01.

Simulation 1										
	A	B	C	D	E	F	G	H	I	J
Median(ρ_p)	0.36	0.78	0.32	0.75	0.06	0.05	0.00	0.48	0.57	0.50
Cluster 1 (5465)	><	<>	00	<>	00	00	00	><	><	<>
Cluster 2 (3159)	<>	><	00	><	00	00	00	><	<>	><
Cluster 3 (1376)	00	><	00	00	00	00	00	<>	<>	<>
Simulation 2										
	A	B	C	D	E	F	G	H	I	J
Median(ρ_p)	0.63	0.38	0.35	0.53	0.00	0.50	0.51	0.16	0.07	0.09
Cluster 1 (1153)	00	00	00	00	00	><	><	00	00	00
Cluster 2 (1926)	<>	><	00	><	00	><	><	00	00	00
Cluster 3 (2031)	<>	<>	><	><	00	><	><	00	00	00
Cluster 4 (2466)	<>	><	<>	<>	00	><	><	00	00	00
Cluster 5 (2424)	><	<>	00	<>	00	<>	<>	00	00	00
Simulation 3										
	A	B	C	D	E	F	G	H	I	J
Median(ρ_p)	0.38	0.50	0.30	0.54	0.07	0.34	0.49	0.41	0.43	0.66
Cluster 1 (7676)	<>	><	00	><	00	><	><	00	00	00
Cluster 2 (2324)	><	<>	00	<>	00	<>	<>	00	00	00
Simulation 4										
	A	B	C	D	E	F				
Median(ρ_p)	0.92	0.87	0.97	0.56	0.70	0.46				
Cluster 1 (2986)	>><	>><	>><	>><	>><	>><				
Cluster 2 (306)	000	<><	<><	0 > 0	< 00	000				
Cluster 3 (700)	>><	>< 0	>><	<><	<><	<><				
Cluster 4 (260)	<><	<><	<><	<><	<><	<><				
Cluster 5 (354)	<><	<><	00 >	000	0 ><	000				
Cluster 6 (394)	<><	0 > 0	<><	000	0 <>	0 ><				
Simulation 5										
	A	B	C	D	E	F	G	H		
Median(ρ_p)	0.96	0.95	0.97	0.93	0.97	0.96	0.97	0.96		
Cluster 1 (4036)	><	<>	<>	<>	<>	><	<>	<>		
Cluster 2 (3813)	><	<>	<>	<>	><	<>	><	><		
Cluster 3 (399)	><	00	<>	><	<>	><	><	<>		
Cluster 4 (720)	<>	><	><	><	<>	><	<>	<>		
Cluster 5 (902)	<>	><	><	><	><	<>	><	><		
Cluster 5 (130)	<>	><	><	><	><	<>	><	<>		
Edwards and Havranek data (CHD)										
	A	B	C	D	E	F				
Median(ρ_p)	0.86	0.92	0.94	0.26	0.81	0.10				
Cluster 1 (900)	><	<>	><	00	<>	00				
Cluster 2 (941)	<>	><	<>	00	><	00				
Genetic-environmental data (GE)										
	rs8034191 (A)	rs4324798 (B)	rs1950081 (C)	age (D)	sex (E)	smoking (F)				
Median(ρ_p)	0.01	0.00	0.10	0.92	0.82	0.85				
Cluster 1 (2222)	00	00	00	><	><	<>				
Cluster 2 (2059)	00	00	00	<>	<>	><				

algorithms to assist log-linear model search algorithms.

$$\boldsymbol{T}_{\gamma}^{sim1} = \begin{pmatrix} & A & B & C & D & E & F & G & H & I & J \\ A & & \mathbf{0.52} & 0.08 & \mathbf{0.50} & 0.04 & 0.02 & 0.02 & 0.20 & 0.27 & 0.15 \\ B & & & \mathbf{0.45} & 1 & 0.06 & 0.04 & 0.03 & 0.47 & 0.64 & 0.47 \\ C & & & & \mathbf{0.45} & 0.02 & 0.02 & 0.009 & 0.12 & 0.23 & 0.16 \\ D & & & & & 0.06 & 0.04 & 0.03 & 0.45 & 0.65 & 0.48 \\ E & & & & & & 0.003 & 0.003 & 0.03 & 0.04 & 0.03 \\ F & & & & & & & 0.002 & 0.02 & 0.03 & 0.03 \\ G & & & & & & & & 0.02 & 0.02 & 0.02 \\ H & & & & & & & & & \mathbf{0.61} & \mathbf{0.56} \\ I & & & & & & & & & & \mathbf{0.74} \end{pmatrix}$$

$$\boldsymbol{T}_{\gamma}^{sim2} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{array} \begin{array}{c} \begin{array}{ccccccccc} A & B & C & D & E & F & G & H & I & J \end{array} \\ \begin{pmatrix} & \mathbf{0.57} & 0.37 & \mathbf{0.72} & 0.04 & \mathbf{0.96} & \mathbf{1} & 0.19 & 0.06 & 0.05 \\ & & \mathbf{0.43} & 0.36 & 0.02 & 0.38 & 0.27 & 0.08 & 0.03 & 0.02 \\ & & & \mathbf{0.63} & 0.02 & 0.24 & 0.24 & 0.05 & 0.03 & 0.03 \\ & & & & 0.03 & 0.48 & 0.54 & 0.13 & 0.04 & 0.05 \\ & & & & & 0.03 & 0.03 & 0.005 & 0.002 & 0.003 \\ & & & & & & \mathbf{0.83} & 0.20 & 0.05 & 0.04 \\ & & & & & & & 0.18 & 0.05 & 0.04 \\ & & & & & & & & 0.01 & 0.01 \\ & & & & & & & & & 0.005 \end{pmatrix} \end{array}$$

$$\boldsymbol{T}_{\gamma}^{sim3} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{array} \begin{array}{c} \begin{array}{ccccccccc} A & B & C & D & E & F & G & H & I & J \end{array} \\ \begin{pmatrix} & \mathbf{0.69} & 0.15 & \mathbf{0.60} & 0.06 & \mathbf{0.48} & \mathbf{0.70} & 0.16 & 0.21 & 0.55 \\ & & \mathbf{0.46} & 0.91 & 0.07 & 0.52 & 0.86 & 0.27 & 0.38 & 0.72 \\ & & & \mathbf{0.69} & 0.02 & 0.10 & 0.27 & 0.14 & 0.19 & 0.30 \\ & & & & 0.07 & 0.34 & 0.70 & 0.27 & 0.39 & 0.66 \\ & & & & & 0.03 & 0.06 & 0.03 & 0.03 & 0.06 \\ & & & & & & \mathbf{0.95} & 0.18 & 0.44 & 0.80 \\ & & & & & & & 0.30 & 0.56 & 1 \\ & & & & & & & & \mathbf{0.49} & \mathbf{0.62} \\ & & & & & & & & & \mathbf{0.81} \end{pmatrix} \end{array},$$

$$\boldsymbol{T}_{\gamma}^{sim4} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} & \mathbf{0.95} & \mathbf{1} & 0.77 & 0.85 & 0.72 \\ & & \mathbf{0.98} & 0.77 & \mathbf{0.83} & 0.72 \\ & & & 0.78 & 0.87 & 0.73 \\ & & & & \mathbf{0.72} & 0.66 \\ & & & & & \mathbf{0.69} \end{pmatrix} \end{array}$$

$$\boldsymbol{T}_{\gamma}^{sim5} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{pmatrix} & \mathbf{0.99} & \mathbf{1} & \mathbf{1} & 1 & 1 & 1 & 1 \\ & & 0.97 & 0.99 & 0.99 & 0.99 & 0.99 & 0.99 \\ & & & 0.99 & 0.99 & 0.99 & 0.99 & 0.99 \\ & & & & 0.99 & 1 & 1 & 1 \\ & & & & & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ & & & & & & \mathbf{1} & \mathbf{1} \\ & & & & & & & \mathbf{1} \end{pmatrix} \end{array}.$$

4.3.4. Log-linear model selection with the aid of the clustering output

Due to the relatively small number of subjects in relation to the number of cells in the contingency tables, and the variability inherent in such simulations, posterior model probabilities are not 80%, 10% and 10% for Models 1,2, and 3 shown in Fig. 1. In Fig. 2, the top 3 models a posteriori as well as model probabilities are presented for each simulation. For simulations 1 to 4, the most likely model a posteriori is the same as the main model used to create the data (Model 1 in Fig. 1), whilst this is not the case for simulation 5. Model probabilities were derived using the Reversible jump algorithm and search strategy (d); see results presented in Table 5.

Simulations 1 to 3 generate contingency tables that are not sparse. The Reversible Jump algorithm can explore the whole set of possible graphical models without removing any covariates from the analysis. In contrast, with regard to simulation 4, the removal of 14 marginally independent covariates reduced the size of the contingency table from 3.4×10^9 to 729 cells, and the number of log-linear graphical models from 1.5×10^{57} to a more manageable 32768. We performed model comparison on the reduced data set with six covariates, using variation (a) where all proposed moves are random, in effect a variation that corresponds to using PDV after reducing the model space with the cluster specific approach. We also consider the three model search variations that utilize $\boldsymbol{T}_{\gamma}$, (b), (c) and (d).

Removing 92 marginally independent covariates from the simulation 5 analysis reduced the size of the contingency table from 1.27×10^{30} to 256 cells, and the number of log-linear graphical models from 2^{4950} to 2^8; a huge gain. Although simulation 5 mainly illustrates the utility of clustering output in reducing the number of covariates for sparse contingency tables, it also illustrates the fact that the converse of Theorem 1 does not hold. For the covariates kept in the analysis, all weights in the $\boldsymbol{T}_{\gamma}$ matrix are effectively equal to one, even for non-existing edges. Consequently, after removing the unimportant covariates, it is not possible to improve on the standard search algorithm by considering the cluster specific output. In fact, model comparison on the reduced data set was performed using only one search strategy, as all four strategies are equivalent. In general, if there is little variability in the elements of the $\boldsymbol{T}_{\gamma}$ matrix, we do not expect that this matrix will be informative to the model search.

Table 5
Mixing performance of samplers. Median of iterations to best model is calculated after 30 runs of the reversible jump MCMC chain. First and third quartiles are given in parentheses. PDV denotes the unrefined model search strategy adopted in Papathomas et al. (2011a). See Fig. 2 for the highest posterior probability model.

	Acceptance rate as a percentage	Iterations (median) to highest posterior probability model	Posterior probability for highest probability model
Simulation 1			
(a) Uniformly random (PDV)	5.1	590 (452,821)	0.55
(b) Cluster specific	3.8	247 (164,369)	0.55
(c) Combined (30%,10%)	5.3	540 (290,674)	0.53
(d) Combined (20%,20%)	4.9	403 (312,493)	0.55
Simulation 2			
(a) Uniformly random (PDV)	4.4	717 (475,990)	0.60
(b) Cluster specific	4.4	189 (147,238)	0.58
(c) Combined (30%,10%)	4.4	417 (346,354)	0.60
(d) Combined (20%,20%)	4.5	257 (181,314)	0.59
Simulation 3			
(a) Uniformly random (PDV)	3.2	657 (545,1065)	0.62
(b) Cluster specific	3.1	445 (335,592)	0.60
(c) Combined (30%,10%)	3.3	538 (431,701)	0.60
(d) Combined (20%,20%)	3.2	560 (368,815)	0.61
Simulation 4 (considering only the 6 important covariates)			
(a) Uniformly random	2.2	661 (550,746)	0.55
(b) Cluster specific	2.08	685 (534,1015)	0.49
(c) Combined (30%,10%)	2.5	625 (543,806)	0.42
(d) Combined (20%,20%)	2.2	733 (551,947)	0.62
Simulation 5 (considering only the 8 important covariates)			
Any of the 4 equivalent strategies	1.1	5183 (3711,6590)	0.74

In Table 5, we present results on the performance of the different reversible jump chains and search strategies. The cluster specific approach (b) outperforms the other search strategies, in terms of iterations to best model. This effect is more prominent in simulations 1 and 2. Search strategy (b) offers a noticeably lower acceptance rate in simulation 1, where we observe a trade-off between acceptance rate and number of iterations to the best model. Intuitively, by having more targeted moves, the overall chance of jumping decreases, but the chain moves more quickly to the higher posterior probability region.

Overall, results in simulations 1 to 3 show the benefit of search strategy (b), where information from variable selection within clustering is included in log-linear model search. With regard to simulation 4, there is little improvement when the $\boldsymbol{T}_\gamma^{sim4}$ matrix is employed; see Table 5. This was expected, as there is little variability in the elements of $\boldsymbol{T}_\gamma^{sim4}$. In the Supplemental material, Section S5 (Appendix B), we examine the rate of accumulated mass of posterior model probability for the first 3 simulations and the different search strategies employed. The reported results also support the argument for incorporating information from variable selection within clustering.

Although our experimental results support search strategy (b), strategy (d), where (a) is combined in a balanced manner with (b), also performs well, offering a good balance between acceptance rate and iterations to best model. Although we did not observe this in any of our analyses, it is prudent to include random search steps that do not depend on the derived $\boldsymbol{T}_\gamma$ matrix as a safeguard, in case variable selection within clustering does not detect an existing edge in a high probability graphical model. In this hypothetical scenario, the search moves that do not depend on $\boldsymbol{T}_\gamma$ will allow for the detection of the covariate space that is not supported by the clustering. Note that edges not reflected in $\boldsymbol{T}_\gamma$ are likely to exist in lower probability models.

5. Real data illustrations

MCMC specifications for the two real data illustrations, as well as run times, are given in Table 3. Prior distributions were the same as the ones adopted in the analysis of the simulated data, described in Section 4.2.

5.1. Risk factors for coronary heart disease

Edwards and Havránek (1985) presented a 2^6 contingency table in which 1841 men were cross-classified by six risk factors for coronary heart disease (CHD). We assume that main effects are always present and compare the 32768 possible graphical log-linear models. Due to the large number of times this data set has been analysed in the past [see, for example, Dellaportas and Forster (1999)] the top two graphical models ('ADE + AC + BC + BE + F' and 'AE + DE + AC + BC + BE + F', following the notation in Agresti (2002)) and associated posterior model probabilities (0.28 and 0.23 respectively for unit information priors) are known. All other graphical models have posterior probabilities lower than 0.1.

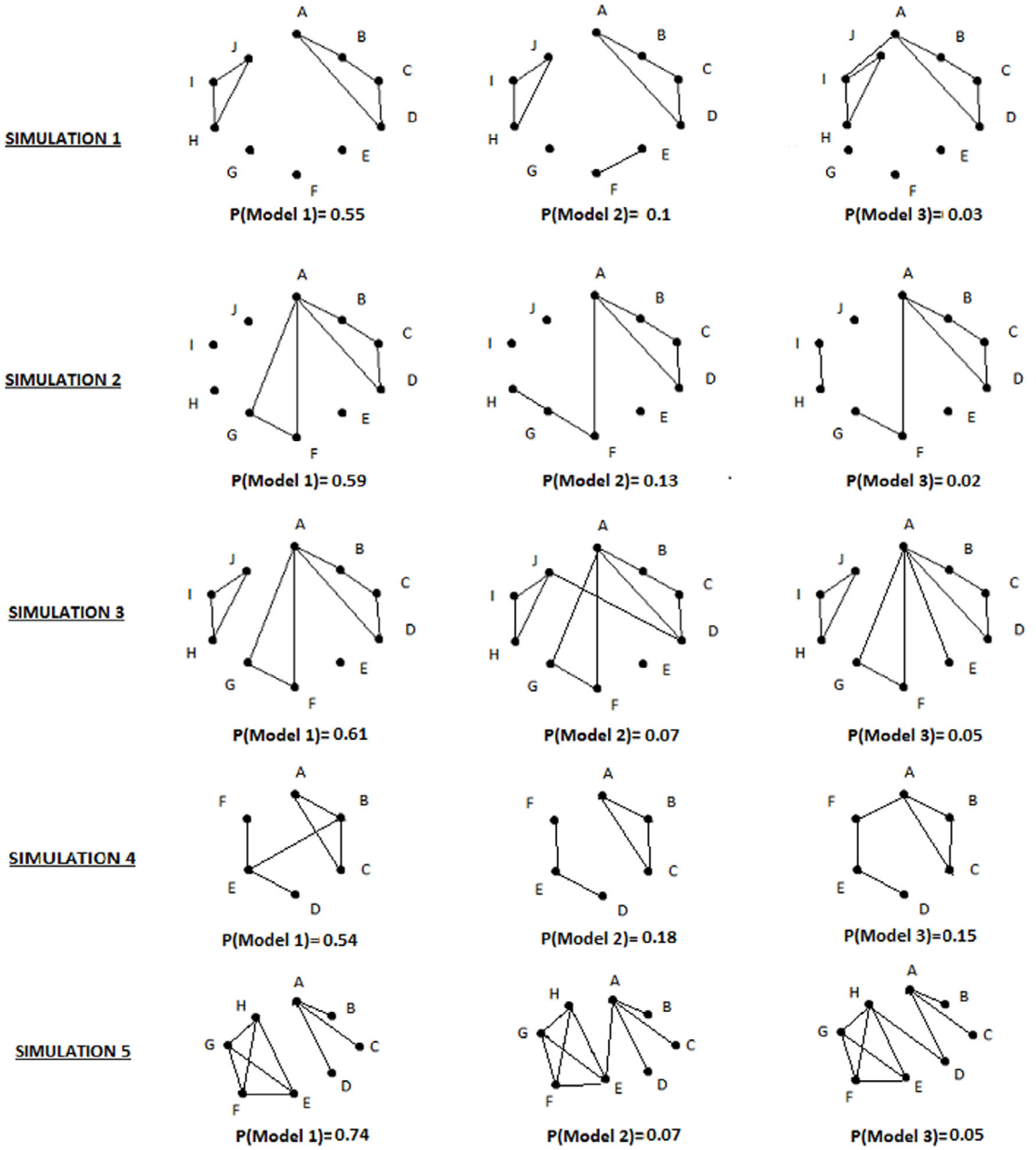

Fig. 2. The resulting best models from the five simulations.

In Table 4, we present the covariate profiles of the representative clusters created with the Bayesian partitioning analysis. The subjects are divided in two clusters, and it is not straightforward to disentangle the log-linear model interactions that are present from the cluster profiles.

The two-way interactions 'AC', 'AE', 'BC' and 'BE' are clearly captured by $\boldsymbol{T}_\gamma$; see below. As in Section 4.3.3, we display with bold font the values of elements that correspond to an existing edge in the most probable model. This demonstrates the applicability of our approach. Elements $t_\gamma(1,4) = 0.14$ and $t_\gamma(4,5) = 0.12$ that correspond to the three-way interaction 'ADE' are smaller. We believe this is due to the signal in the data not being strong. The two likely models have combined posterior probability equal to 0.51, whilst only one of them contains the three-way interaction 'ADE'. No other model is associated with probability greater than 0.1. Nevertheless, the two elements $t_\gamma(1,4)$ and $t_\gamma(4,5)$ are still one order of magnitude larger compared to the five elements that correspond to 'F'. Factor 'F' does not interact with any other covariate, and this matches the low posterior selection probability $E(\rho_6) = 0.10$, implying it is not likely to propose the addition of an edge in the graphical model from covariate 'F' to another covariate. Of the eleven edges that are not present in the high

Table 6
Mixing performance of samplers. Median of iterations to best model is calculated after 300 runs of the reversible jump MCMC chain. First and third quartiles are given in parentheses. PDV denotes the unrefined model search strategy adopted in Papathomas et al. (2011a).

Edwards and Havranek data (CHD)	Acceptance rate as a percentage	Iterations (median) to highest posterior probability model	Posterior probability for highest probability model 'ADE + AC + BC + BE + F'
(a) Uniformly random (PDV)	5.2	314 (215,582)	0.28
(b) Cluster specific	3.7	244 (162,378)	0.28
(c) Combined (30%,10%)	4.9	273 (172,470)	0.27
(d) Combined (20%,20%)	4.6	248 (155,392)	0.28
Genetic-environmental data [including important (characterized as such by clustering) representative SNPs]	Acceptance rate as a percentage	Iterations (median) to highest posterior probability model	Posterior probability for highest probability model 'A + B + C + DEF'
(a) Uniformly random	6.3	564 (257,1205)	0.53
(b) Cluster specific	8.4	196 (83,443)	0.51
(c) Combined (30%,10%)	6.9	310 (147,670)	0.51
(d) Combined (20%,20%)	7.5	235 (91,516)	0.52

probability model, five correspond to very small elements t_γ. Using $\boldsymbol{T}_\gamma$ to inform the model search algorithm, results in the identification of a large part of the model space that is associated with low probability.

$$\boldsymbol{T}_\gamma^{\text{Real data (CHD)}} = \begin{pmatrix} & A & B & C & D & E & F \\ A & & 0.81 & \mathbf{0.81} & \mathbf{0.14} & \mathbf{0.56} & 0.04 \\ B & & & \mathbf{1} & 0.16 & \mathbf{0.75} & 0.05 \\ C & & & & 0.16 & 0.75 & 0.05 \\ D & & & & & \mathbf{0.12} & 0.01 \\ E & & & & & & 0.05 \end{pmatrix}.$$

In Table 6, we present model selection results. It is clear that adopting search strategy (b) to incorporate information from the clustering analysis reduces the average number of iterations to the best model. Model search strategy (d), where (a) and (b) are combined also performs well, as was the case in the simulations.

5.2. Genetic and other risk factors

We consider thirty single nucleotide polymorphisms (SNPs) in chromosomes 6 and 15. These are data from 4260 subjects that participated in a genome-wide association study of lung cancer presented in Hung et al. (2008). The thirty most significant SNPs in terms of marginal p-value are analysed. Some of these genetic markers were identified as associated with the phenotype in Papathomas et al. (2012). We consider two levels for each marker (0 – wild type; 1 – homozygous or heterozygous variant).

Twelve SNPs were indicated as important by variable selection within clustering; two from chromosome 15 and ten from chromosome 6. Nine of the selected chromosome 6 SNPs are highly correlated. The two selected chromosome 15 SNPs are also highly correlated. Therefore, we decided to include three SNPs in the log-linear graphical model as representatives of the selected SNPs; rs8034191 from chromosome 15 and {rs4324798,rs1950081} from chromosome 6. We also include age, gender and smoking status in the log-linear graphical model, to search for gene-environment interactions as well as gene–gene interactions. We consider two levels for smoking (0 – non or ex smoker; 1 – smoker) and age (below and above median). The variables will be referred to as A to F, with {A,B,C} denoting the genetic factors.

Reducing the number of SNPs from 30 to 12, and then to 3, allows for the use of reversible jump MCMC to compare competing graphical models. The 2^{33} contingency table would be too sparse with the vast majority of cells equal to zero.

The highest posterior probability model is 'A + B + C + DEF', which does not support the presence of gene–gene or gene-environment interactions. On the other hand, a three-way interaction 'DEF' is suggested, which implies different patterns of smoking behaviour by age and gender. The presence of such an interaction is in line with epidemiological understanding, and shows that our algorithm performs well. In Table 4, we present the profiles of the representative clusters created with the Bayesian partitioning analysis. The subjects are typically divided in two clusters which correspond to the 'DEF' interaction.

The derived $\boldsymbol{T}_\gamma$ matrix is shown below, after the first stage clustering analysis is performed afresh for the six covariates. We did not cluster the subjects using all 12+3 covariates because the 12 highly correlated important SNPs would 'swamp' the 3 environmental factors. The $\boldsymbol{T}_\gamma$ matrix correctly indicates the presence of the three-way interaction 'DEF'. It also correctly indicates that the first three covariates do not form any interaction terms. In this case, we do see a close correspondence

between the clustering pattern and interactions in the associated log-linear model.

$$\boldsymbol{T}_{\gamma}^{\text{Real data (GE) (2nd run)}} = \begin{pmatrix} & A & B & C & D & E & F \\ A & & 0.002 & 0.01 & 0.06 & 0.06 & 0.06 \\ B & & & 0.001 & 0.02 & 0.02 & 0.02 \\ C & & & & 0.09 & 0.07 & 0.08 \\ D & & & & & \mathbf{1} & \mathbf{0.98} \\ E & & & & & & \mathbf{0.88} \end{pmatrix}.$$

Similarly to the previous real data analysis, using the $\boldsymbol{T}_{\gamma}$ matrix to inform the model search algorithm results in the identification of part of the model space that is associated with low probability and improvement in model search (Table 6).

We also investigated an alternative approach for assessing the evidence for the presence of an edge, where the pairwise association between two factors is evaluated by the estimation of odds-ratios. See the Supplemental material, Section S6 (Appendix B), for more details on these calculations, as well as an illustration on the two real data sets analysed in this manuscript. Results demonstrate that our approach, based on a clustering procedure that considers all variables simultaneously, gives different information on the presence of interactions (two-way and higher) than an approach which is based purely on pairwise associations. For example, for the genetic data analysed in this subsection, the pairwise approach fails to capture an association between D and F, despite the three-way interaction 'DEF' present in the prominent highest posterior probability model; see Table S2 in the Supplemental material (Appendix B). We further discuss this in the next section.

6. Discussion

The advantage in utilizing variable selection within partitioning to inform log-linear model selection is mostly pertinent to marginal independence. For sparse contingency tables, this information can lead to the substantial reduction of the number of covariates considered, making the exploration of the model space feasible. For example, in the second real data illustration, it would be impossible to explore the model space for a 2^{33} contingency table by conventional methods such as the Reversible jump MCMC, without the considerable reduction in the number of SNPs through the first clustering stage. Theoretical results presented in Section 3.1 show that covariates $x_{\cdot p}$ with posterior median selection probability ρ_p equal to zero (or very close to zero in practice) do not form interaction terms. This appears to be true even when a sparse prior distribution is adopted for the selection parameters ρ_p, as was the case for all simulation studies and real data analyses in this manuscript.

With regard to detecting conditional independence, utilizing the output from a clustering model, where all variables are considered simultaneously, offers different results compared to methods based on pairwise associations for the detection of edges. This was illustrated empirically on the two real data examples; see results presented in the Supplemental material (Appendix B). Intuitively, considering all variables simultaneously, rather than in a pairwise fashion, should increase the likelihood of detecting dependence structures that are more complex than pairwise dependencies such as two-way interactions. Nonetheless, it is possible that incorporating in some manner information coming from odds-ratios could be beneficial, given that multiple testing concerns are addressed. Note that our approach utilizes a variable selection approach where all factors are included simultaneously in the model, with a prior assigned to the probability of inclusion. This makes it less susceptible to multiple testing concerns, and particularly suitable for reducing the search space in cases where a large number of factors is investigated; see Scott and Berger (2010).

Adopting search strategy (b) and informing the model search algorithm with $\boldsymbol{T}_{\gamma}$ often improves the efficiency of the search. Although marginal independence was not always detected, because the converse of the Theorems in 3.1 does not hold, in the majority of the analyses $\boldsymbol{T}_{\gamma}$ identified parts of the model space that contained models of low probability, leading to more efficient model search steps. Importantly, using $\boldsymbol{T}_{\gamma}$ to assist the model search never resulted in a worse algorithm, compared to the standard model search approach in Papathomas et al. (2011a). In terms of number of iterations to the best model, the model search algorithm that is informed by clustering performed better or at least as efficiently as the standard algorithm. The additional computational cost for the clustering is minimal when the R package PReMiuM is used (Liverani et al., 2015), which is primarily written in C++ and R; see the run times reported in Table 3. The approach where the naive model search (a) is combined in a balanced manner with (b), where the $\boldsymbol{T}_{\gamma}$ matrix is employed, also performs well, offering a good balance between acceptance rate and number of iterations to the best model. Combining a 'naive' with a more 'targeted' search approach ensures a comprehensive and efficient exploration of the model space, in the same spirit as the simultaneous sampling from 'hot' and 'cold' chains in simulated tempering (Geyer and Thompson, 1995).

In Johndrow et al. (2014), the authors consider standard and novel latent class structures. The DP is a special case, and its rank is defined as the minimum number of clusters required to describe the joint probability tensor for the categorical covariates. The authors relate log-linear modelling with latent class modelling, investigating if a trivial relationship exists between the two modelling approaches, as we do in this manuscript, albeit from a different standpoint. Bounds are derived for the rank of the latent class model, in relation to the number and structure of the interactions that are present in a weakly hierarchical log-linear model. In one of the results, a massive reduction in the upper bound of the latent class model's rank is shown, under a sparse log-linear model; a model is defined as sparse when the number of non-zero model parameters is much smaller compared to the number of parameters in the saturated model. The authors also demonstrate that the rank

of the latent structure depends only on variables that are not marginally independent. A straightforward application of one of the results in Johndrow et al. (2014), gives that an upper bound of the rank of the latent class model corresponding to the prominent model of simulation 1 is 2^7, rather than the default 2^9. The upper bound corresponding to the prominent model of simulation 5 is 2^8, rather than the default 2^{99}.

Zhou et al. (2015) also utilizes the idea that marginally independent variables reduce the dimensionality of the model required to describe the joint probability distribution between the covariates. A PARAFAC factorization is adopted, which can be viewed as a more general representation of the Dirichlet process. Dimensionality reduction is achieved with the introduction of the sparse PARAFAC (sp-PARAFAC) formulation, where marginal independence is modelled with fixed baseline vectors, quantities that correspond to the $\pi_p(x)$ quantities we introduced in this manuscript. These are the main similarities between the two approaches, although there are significant differences too. In Zhou et al. (2015) the focus of the theoretical results are in providing expressions for parameters of the log-linear models that correspond to the adopted latent class model, assessing the level of induced shrinkage, and assessing the convergence of the probability tensor induced by the sp-PARAFAC formulation to the true probability tensor. In contrast, we focus our theoretical investigation on the variable selection switches and what they imply with regard to marginal independence. The prior formulation for detecting marginally independent covariates and reducing dimensionality is also different in the two approaches. Finally, the objectives in the two manuscripts are different, as we focus on accelerating log-linear model selection with the Reversible Jump approach by utilizing output from the clustering process.

A limitation of the approach introduced in this manuscript, as well other approaches we discussed, is the inability to detect conditional independence through the clustering output in a consistent and wieldy manner. One recent attempt at tackling this problem is Kunihama and Dunson (2014), where the concept of mutual information is introduced. Results similar to the ones in Section 3.1, concerning conditional independence, would be useful as conditional independence between variables is key when building the joint distribution of $\{x_{\cdot 1}, \ldots, x_{\cdot P}\}$ using graphical models. Investigating a possible direct link between variable selection within clustering and conditional independence is the subject of ongoing research.

Acknowledgements

We would like to thank the Editor, Associate Editor and three reviewers for comments that improved this manuscript. We would also like to thank Professor Paolo Vineis and Dr Paul Brennan for providing the data used in Section 5.2. This work was supported by MRC grant G1002319.

Appendix A

Proof of Theorem 1. Assume that the subjects are grouped into C clusters. As $\sum_{c=1}^{C} \gamma_p^c \times \gamma_q^c = 0$, without any loss of generality, assume that, for $x_{\cdot p}$ and $x_{\cdot q}$,

$$\begin{aligned}
&\gamma_p^c = 0, \gamma_q^c = 1, \quad \text{for } c \in \Gamma_1,\\
&\gamma_p^c = 1, \gamma_q^c = 0, \quad \text{for } c \in \Gamma_2,\\
&\gamma_p^c = 0, \gamma_q^c = 0, \quad \text{for } c \in \Gamma_3 = \{1, \ldots, C\} \cap (\Gamma_1 \cup \Gamma_2)^{\complement},
\end{aligned}$$

where $\Gamma_1 \cap \Gamma_2 = \emptyset$. To simplify the notation, we suppress the x and x' from $P(x_{\cdot p} = x, x_{\cdot q} = x')$, and write $P(x_{\cdot p}, x_{\cdot q})$. We also write ϕ_p^c instead of $\phi_p^c(x)$, and π_p instead of $\pi_p(x)$. Finally, we write $\sum_{\Gamma_l}$, $l = 1, 2, 3$, instead of $\sum_{c \in \Gamma_l}$. Then,

$$\begin{aligned}
P(x_{\cdot p}, x_{\cdot q}) &= \sum_{c=1}^{C} \psi_c \{(\phi_p^c)^{\gamma_p^c} (\pi_p)^{1-\gamma_p^c}\} \{(\phi_q^c)^{\gamma_q^c} (\pi_q)^{1-\gamma_q^c}\}\\
&= \pi_p \sum_{\Gamma_1} \psi_c \phi_q^c + \pi_q \sum_{\Gamma_2} \psi_c \phi_p^c + \pi_p \pi_q \sum_{\Gamma_3} \psi_c.
\end{aligned}$$

Also,

$$\begin{aligned}
P(x_{\cdot p})P(x_{\cdot q}) &= \left(\sum_{\Gamma_1} \psi_c \pi_p + \sum_{\Gamma_2} \psi_c \phi_p^c + \sum_{\Gamma_3} \psi_c \pi_p\right) \times \left(\sum_{\Gamma_1} \psi_c \phi_q^c + \sum_{\Gamma_2} \psi_c \pi_q + \sum_{\Gamma_3} \psi_c \pi_q\right)\\
&= \pi_p \left(\sum_{\Gamma_1} \psi_c \phi_q^c\right) \left(1 - \sum_{\Gamma_2} \psi_c\right) + \pi_q \left(\sum_{\Gamma_2} \psi_c \phi_p^c\right) \left(1 - \sum_{\Gamma_1} \psi_c\right)\\
&\quad + \pi_p \pi_q \left\{\left(\sum_{\Gamma_1} \psi_c\right) \left(\sum_{\Gamma_2} \psi_c\right) + \left(\sum_{\Gamma_3} \psi_c\right)\right\} + \left(\sum_{\Gamma_2} \psi_c \phi_p^c\right) \left(\sum_{\Gamma_1} \psi_c \phi_q^c\right).
\end{aligned}$$

Now,

$$
\begin{aligned}
&P(x_{\cdot p}, x_{\cdot q}) - P(x_{\cdot p})P(x_{\cdot q}) = 0 \\
&\Leftrightarrow \pi_p \left(\sum_{\Gamma_1} \psi_c \phi_q^c\right)\left(\sum_{\Gamma_2} \psi_c\right) + \pi_q \left(\sum_{\Gamma_2} \psi_c \phi_p^c\right)\left(\sum_{\Gamma_1} \psi_c\right) \\
&\quad - \pi_p \pi_q \left\{\left(\sum_{\Gamma_1} \psi_c\right)\left(\sum_{\Gamma_2} \psi_c\right)\right\} - \left(\sum_{\Gamma_2} \psi_c \phi_p^c\right)\left(\sum_{\Gamma_1} \psi_c \phi_q^c\right) = 0 \\
&\Leftrightarrow \left\{\pi_p \left(\sum_{\Gamma_2} \psi_c\right) - \sum_{\Gamma_2} \psi_c \phi_p^c\right\}\left\{\sum_{\Gamma_1} \psi_c \phi_q^c - \pi_q \left(\sum_{\Gamma_1} \psi_c\right)\right\} = 0
\end{aligned}
$$

This is always true since, for example, as $\pi_p(x) = P(x_{\cdot p} = x)$,

$$
\begin{aligned}
&\pi_p = P(x_{\cdot p}) = \sum_c \psi_c (\phi_p^c)^{\gamma_p^c} (\pi_p)^{1-\gamma_p^c} = \pi_p \sum_{\Gamma_1 \cup \Gamma_3} \psi_c + \sum_{\Gamma_2} \psi_c \phi_p^c \\
&\Rightarrow \sum_{\Gamma_2} \psi_c \phi_p^c = \pi_p - \pi_p \left(1 - \sum_{\Gamma_2} \psi_c\right) \\
&\Rightarrow \sum_{\Gamma_2} \psi_c \phi_p^c = \pi_p \sum_{\Gamma_2} \psi_c .
\end{aligned}
$$

Proof of Theorem 2. Without loss of generality, to simplify the notation assume that $p = 1$. Then, for all $q \in \{2, \ldots, \ldots, P\}$, $\sum_{c=1}^{C} \gamma_1^c \times \gamma_q^c = 0$. From Theorem 1, $x_{\cdot 1}$ is independent of $x_{\cdot q}$, for any $2 \le q \le P$. Such pairwise independence does not imply that $x_{\cdot 1}$ is independent of $\{x_{\cdot 2}, \ldots, x_{\cdot P}\}$. To show this assume, also without loss of generality, that $\gamma_1^c = 0$, for $c \in \Gamma_1$ and $\gamma_1^c = 1$, for $c \in \Gamma_2$. The Γ_1 and Γ_2 sets can be empty. Now, since, $\sum_{c=1}^{C} \gamma_1^c \times \gamma_q^c = 0$, for all $q \in \{2, \ldots, \ldots, P\}$, $\gamma_q^c = 0$ for all $c \in \Gamma_2$. Then,

$$
\begin{aligned}
P(x_{\cdot 1}, x_{\cdot 2}, \ldots, x_{\cdot P}) &= \sum_{c=1}^{C} \psi_c \{(\phi_1^c)^{\gamma_1^c} (\pi_1)^{1-\gamma_1^c}\} \times \{(\phi_2^c)^{\gamma_2^c} (\pi_2)^{1-\gamma_2^c}\} \times \cdots \times \{(\phi_P^c)^{\gamma_P^c} (\pi_P)^{1-\gamma_P^c}\} \\
&= \pi_1 \times \sum_{\Gamma_1} \psi_c \{(\phi_2^c)^{\gamma_2^c} (\pi_2)^{1-\gamma_2^c}\} \times \cdots \times \{(\phi_P^c)^{\gamma_P^c} (\pi_P)^{1-\gamma_P^c}\} + \pi_2 \times \cdots \times \pi_P \times \sum_{\Gamma_2} \psi_c \phi_1^c .
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\pi_1 = \sum_{c \in \Gamma_1 \cup \Gamma_2} \psi_c (\phi_1^c)^{\gamma_1^c} (\pi_1)^{1-\gamma_1^c} = \pi_1 \sum_{\Gamma_1} \psi_c + \sum_{\Gamma_2} \psi_c \phi_1^c \\
&\Rightarrow \sum_{\Gamma_2} \psi_c \phi_1^c = \pi_1 - \pi_1 \left(1 - \sum_{\Gamma_2} \psi_c\right) \\
&\Rightarrow \sum_{\Gamma_2} \psi_c \phi_1^c = \pi_1 \sum_{\Gamma_2} \psi_c .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P(x_{\cdot 1}, x_{\cdot 2}, \ldots, x_{\cdot P}) \\
&= \pi_1 \left(\sum_{\Gamma_1} \psi_c \{(\phi_2^c)^{\gamma_2^c} (\pi_2)^{1-\gamma_2^c}\} \times \cdots \times \{(\phi_P^c)^{\gamma_P^c} (\pi_P)^{1-\gamma_P^c}\} + \pi_2 \times \cdots \times \pi_P \times \sum_{\Gamma_2} \psi_c\right) \\
&= P(x_{\cdot 1}) \times P(x_{\cdot 2}, \ldots, x_{\cdot P}),
\end{aligned}
$$

and $x_{\cdot 1}$ is independent of $\{x_{\cdot 2}, \ldots, x_{\cdot P}\}$ as required.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jspi.2016.01.002.

References

Agresti, A., 2002. Categorical Data Analysis, second ed.. John wiley & Sons, New Jersey.
Bhattacharya, A., Dunson, D.B., 2012. Simplex factor models for multivariate unordered categorical data. J. Amer. Statist. Assoc. 107, 362–377.
Bingham, S., Riboli, E., 2004. Diet and cancer — the European prospective Investigation into cancer and nutrition. Nature Rev. Cancer 4, 206–215.
Burton, P.R., Hansell, A.L., Fortier, I., Manolio, T.A., Khoury, M.J., Little, J., Elliot, P., 2009. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. Int. J. Epidemiol. 38, 263–273.
Cho, H., Fryzlewicz, P., 2012. High dimensional variable selection via tilting. J. R. Stat. Soc. Ser. B 74, 593–622.
Chung, Y., Dunson, D.B., 2009. Nonparametric Bayes conditional distribution modelling with variable selection. J. Amer. Statist. Assoc. 104, 1646–1660.
Clyde, M., George, E.I., 2004. Model uncertainty. Statist. Sci. 19, 81–94.
Dellaportas, P., Forster, J.J., 1999. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86, 615–633.
Dobra, A., 2009. Variable selection and dependency networks for genomewide data. Biostatistics 10, 621–639.
Dobra, A., Massam, H., 2010. The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Stat. Methodol. 7, 240–253.
Dunson, D.B., Herring, A.H., Engel, S.M., 2008. Bayesian selection and clustering of polymorphisms in functionally-related genes. J. Amer. Statist. Assoc. 103, 534–546.
Dunson, D.B., Xing, C., 2009. Nonparametric Bayes modelling of multivariate categorical data. J. Amer. Statist. Assoc. 104, 1042–1051.
Edwards, D., Havránek, T., 1985. A fast procedure for model search in multi-dimensional contingency tables. Biometrika 72, 339–351.
Ferguson, T.S., 1973. A Bayesian analysis of nonparametric problems. Ann. Statist. 1, 209–230.
Forster, J., Gill, R., Overstall, A., 2012. Reversible jump methods for generalised linear models and generalised linear mixed models. Statist. Comput. 22, 107–120.
Geyer, C.J., Thompson, E.A., 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. J. Amer. Statist. Assoc. 90, 909–920.
Green, P.J., 1995. Reversible jump MCMC computation and Bayesian model determination. Biometrika 82, 711–732.
Green, P.J., Richardson, S., 2001. Modelling heterogeneity with and without the Dirichlet process. Scand. J. Stat. 28, 355–375.
Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for 'Large p' regression. J. Amer. Statist. Assoc. 102, 507–516.
Huelsenbeck, J.P., Andolfatto, P., 2007. Inference of population structure under a Dirichlet process model. Genetics 175, 1787–1802.
Hung, R.J., McKay, J.D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., et al., 2008. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452, 633–637.
Ishwaran, H., James, L., 2001. Gibbs sampling methods for stick-breaking priors. J. Amer. Statist. Assoc. 96, 161–173.
Johndrow, J.E., Bhattacharya, A., Dunson, D.B., 2014. Tensor decompositions and sparse log-linear models. arXiv:1404.0396v1.
Kunihama, T., Dunson, D., 2014. Nonparametric Bayes inference on conditional independence. arXiv:1404.1429v1.
Lauritzen, S.L., 2011. Elements of graphical models. In: Lectures from the XXXVIth International Probability Summer School in St-Flour, France. http://www.stats.ox.ac.uk/steffen.
Liverani, S., Hastie, D.I., Azizi, L., Papathomas, M., Richardson, S., 2015. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. J. Statist. Softw. 64 (7), 1–30.
Lo, A.Y., 1984. On a class of Bayesian nonparametric estimates. I. Density estimates. Ann. Statist. 12, 351–357.
MacEachern, S.N., Müller, P., 1998. Estimating mixture of Dirichlet process models. J. Comput. Graph. Statist. 7, 223–238.
Marbac, M., Biernacki, C., Vandewalle, V., 2014. Model-based clustering for conditionally correlated categorical data. arXiv:1401.5684v2.
Molitor, J., Papathomas, M., Jerrett, M., Richardson, S., 2010. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics 11, 484–498.
Ntzoufras, I., Dellaportas, P., Forster, J.J., 2003. Bayesian variable and link determination for generalized linear models. J. Statist. Plann. Inference 111, 165–180.
Papathomas, M., 2015. On the correspondence between Bayesian log-linear and logistic regression models with g-priors, http://arxiv.org/abs/1409.3795.
Papathomas, M., Dellaportas, P., Vasdekis, V.G.S., 2011a. A novel reversible jump algorithm for generalized linear models. Biometrika 98, 231–236.
Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., Richardson, S., 2012. Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene-gene patterns. Genet. Epidemiol. 36, 663–674.
Papathomas, M., Molitor, J., Riboli, E., Richardson, S., Vineis, P., 2011b. Examining the joint effect of multiple risk factors using exposure risk profiles: Lung cancer in non-smokers. Environ. Health Perspect. 119, 84–91.
Reich, B.J., Bondell, H.D., 2011. A spatial Dirichlet process mixture model for clustering population genetics data. Biometrics 67, 381–390.
Richardson, S., Bottolo, L., Rosenthal, J.S., 2010. Bayesian models for sparse regression analysis of high dimensional data. Bayesian Stat. 9, 539–569.
Scott, J.G., Berger, J.O., 2010. Bayes and Empirical Bayes multiplicity adjustment in the variable selection problem. Ann. Statist. 38, 2587–2619.
Sinha, S., Mallick, B.K., Kipnis, V., Carroll, R.J., 2010. Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. Biometrics 66, 444–454.
Wakefield, J., De Vocht, F., Hung, R.J., 2010. Bayesian mixture modelling of gene-environment and gene-gene interactions. Genet. Epidemiol. 34, 16–25.
Walker, S., Damien, P., Laud, P., Smith, A., 1999. Bayesian nonparametric inference for random distributions and related functions (with discussion). J. R. Stat. Soc. Ser. B 61, 485–527.
West, M., 1992. Hyperparameter Estimation in Dirichlet Process Mixture Models. Institute of Statistics and Decision Sciences.
Zhang, W., Zhu, J., Schadt, E.E., Liu, J.S., 2010. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. PLoS Comput. Biol. 6, 1–10.
Zhou, J., Bhattacharya, A., Herring, A.H., Dunson, D.B., 2015. Bayesian factorizations of big sparse tensors. J. Amer. Statist. Assoc. 110, 1562–1576.